Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings

Eugenia San Segundo & Pedro Gómez-Vilda

Consejo Superior de Investigaciones Científicas,
Universidad Internacional Menéndez Pelayo
&
Center for Biomedical Technology,
Universidad Politécnica de Madrid

Abstract. *In this study we have analyzed 853 tokens of the vowel filler [e:], extracted from spontaneous speech fragments of 54 male Spanish speakers (North-Central Peninsular variety), each one recorded on two separate sessions. The speakers — to be compared in a pairwise fashion – were divided in four groups: 24 monozygotic (MZ) twins, 10 dizygotic (DZ) twins, 8 non-twin brothers and 12 unrelated speakers. From the extracted vowel fillers, considered long enough for a glottal analysis (around 160 milliseconds), a vector of 68 glottal parameters was created. Our hypothesis that higher similarity values would be found in the intra-pair comparison of MZ twins than in DZ twins, brothers or unrelated speakers was confirmed, which suggests that the glottal parameters under investigation are genetically influenced. This finding seems of great forensic importance, as a phonetic parameter is considered forensically robust provided that it exhibits large between-speaker variation while it remains as consistent as possible for each speaker (i.e. small within-speaker variation).*

Resumo. *Neste trabalho foram analisadas 853 amostras de preenchimento da vogal [e:], extraídas a partir de fragmentos de fala espontânea de 54 falantes espanhóis do sexo masculino (variedade de fala Norte-Central Peninsular), cada um gravado em duas sessões separadas. Os falantes — comparados dois a dois — foram divididos em quatro grupos: 24 gêmeos monozigóticos (MZ), 10 gêmeos dizigóticos (DZ), 8 irmãos não gêmeos e 12 falantes sem parentesco. A partir das vogais de preenchimento extraídas, consideradas suficientemente longas para uma análise glotal (cerca de 160 milissegundos), um vector de 68 parâmetros glotais foi criado. Nossa hipótese de que seriam encontrados valores de similaridade mais elevados na comparação intra-par dos gêmeos monozigóticos do que na dos*

gêmeos DZ, dos irmãos ou dos falantes sem parentesco foi confirmada, o que sugere que os parâmetros glotais sob investigação são geneticamente influenciados. Essa descoberta parece ser de grande importância forense, na medida em que um parâmetro fonético é considerado robusto para a área forense desde que contenha uma grande variação entre-falantes, enquanto permanece tão consistente quanto possível para cada falante (ou seja, pequena variação intra-falante).

Introduction

In this investigation we have explored the voice characteristics of four speaker groups: monozygotic (MZ) twins, dizygotic (DZ) twins, male non-twin siblings (i.e. brothers, B) and unrelated speakers (US). Among other possible phonetic parameters that could be analyzed in these speakers with forensic purposes (see San Segundo, 2014), on this occasion we have focused on a group of glottal features reported to show good identification results in previous studies (Gómez-Vilda *et al.*, 2010, 2012).

In this introduction we will first describe — in a succinct way — the scientific field to which this study mostly contributes: Forensic Phonetics, and more specifically Forensic Speaker Comparison (FSC)[1] Secondly, we will explain the relevance of the *twin methodology* for this discipline. In a third stage, we will specifically detail how glottal source features have proved useful to discriminate speakers in several studies. This will serve as a state-of-the-art background against which the research hypothesis can be set, together with the methodology, in the next section.

Forensic Phonetics is the application of Phonetics aimed at solving any type of legal issue, or, in the words of Jessen (2008: 671), "the application of the knowledge, theories and methods of general phonetics to practical tasks that arise out of a context of police work or the presentation of evidence in court". There are many tasks which a phonetician may be requested to perform for forensic purposes. French (1994), Rose (2002) and French and Stevens (2013) are only some references where all these forensic tasks are explained in some detail. A brief overview of task classifications by the above-mentioned authors can be read in San Segundo (2014), where five task subgroups are mentioned: (1) determination of unclear or contested utterances — closely related to phonetic transcription; (2) examination of the authenticity of audio recordings; (3) design and validation of voice line-ups; (4) speaker profiling and LADO (Language Analysis for the Determination of Origin of Asylum Seekers)[2]; and (5) Forensic Speaker Comparison (FSC from now on). Out of all these tasks, the one for which speech experts are more frequently required, according to French and Stevens (2013), is the last one. In such cases, the experts have to compare the voice of an *offender* (i.e. the speech samples of an unknown speaker) with the voice of a suspect or several suspects (i.e. the speech samples of known origin). The kind of criminal offenses which are typically involved in FSC usually take place over the telephone, whether they are cases of drug dealers arranging illegal transactions, fraudulent bank deals, bomb hoaxes, kidnappers' ransom demands, or stalking offenses.

Twin studies are not especially widespread in Forensic Phonetics, despite the fact that other forensic disciplines have evinced a clear interest in twin discriminability, particularly in recent times. In San Segundo (2013), some examples were mentioned belonging to DNA testing, fingerprint identification and handwriting discrimination of

twins. A more extensive review of voice-related studies focusing on twins is provided in San Segundo (2014), where thirty-nine works were described, encompassing the year span 1948-2014. What all of them have in common is that they tackle the issue of voice similarity in twins and non-twin siblings, either from an articulatory, acoustical, perceptual or automatic point of view. Of course, not all of the reviewed studies stem from a forensic-phonetic perspective (e.g. trying to answer research questions of interest for this field) but most of them draw on the twin methodology. In other words, they involve the comparison of monozygotic (MZ) and dizygotic (DZ) twins with the aim of finding the relative contribution of genetic and environmental factors on the differences found between them. The twin research methodology offers several design variations of the classic twin method, which compares the resemblance within MZ twin pairs to the resemblance within DZ twin pairs, assuming equal environment influences for both types of twins:

> Differences within MZ twin pairs are explained by environmental effects because all genetic inheritance is commonly shared. In contrast, differences within DZ twin pairs are associated with both genetic and environmental influences because these twins share half their genes, on average, by descent. (Segal, 1990: 613)

In other words, what the twin methodology suggests is that any 'excess' of similarity in MZs over DZs refers to "the proportion of phenotypic variation that can be attributed to genetic variance" (Tomblin and Buckwalter, 1998: 189). In San Segundo (2014) we described the genetic endowment of MZ and DZ twins (100% shared genes in the former, and 50% shared genes in the latter) as well as the environmental influences possibly affecting their voice and speech. In relation to this last aspect, we tried to link "environmental influences" not only to the prenatal-perinatal-postnatal division provided for instance by Stromswold (2006), but also to sociolinguistic perspectives which provide insightful observations about the effects exerted by the family on the linguistic output of individuals (Hazen, 2002). Equally important in this respect are the existing investigations evolving around the idea of 'intratwin mimetism' (Debruyne *et al.*, 2002), which would be more commonly found in MZ than in DZ twins.

All in all, the forensic importance of investigating twins' voices lies in the fact that these speakers are the most extreme cases of physical similarity in human beings. The fact that they are genetically identical — in the case of MZ twins — or very similar — in the case of DZ twins — and most frequently raised in the same circumstances, make their voices highly confusable. Distinguishing them is therefore a challenge in a forensic context, as acknowledged by authors such as Künzel (2010). Some real cases involving the forensic comparison of speech samples in twins and non-twin siblings can be found in Rose (2002) or Rose (2006). Furthermore, Mora (2013) described in a recent piece of news how the perpetrator of six rapes in France could not be clearly identified on the basis of DNA, resulting in the arrest of two MZ twins. Having acknowledged the existence of real offences involving twins — which suggests that the study of these speakers is not so exotic as one could a priori think — it should be pointed out that there is an interest in this kind of investigations *per se*. As explained in San Segundo (2014: 1), "the study of genetically identical speakers (MZ twins) and their comparison with non-identical siblings [...] allows gaining insight into the contribution of nurture and nature in the speech patterns of speakers in general". See next section for a more in-depth explana-

tion of our main hypothesis: the more genetically influenced a phonetic parameter is, the more robust it will be for general speaker comparison.

A final aspect that we would like to highlight in this introduction is related to the use of glottal features for speaker comparison. Current methodologies in FSC are varied and they imply the analysis of multiple features. Indeed it is not uncommon to characterize this forensic-phonetic subdiscipline by its lack of consensus over the analysis and comparison techniques used, but also over issues like the expression of conclusions. Cambier-Langeveld (2007) and Gold and French (2011) provide good summaries of the most common international practices in FSC, with some detailed information about most frequent acoustic measures, relative weighting attached to those parameters, as well as an attempt to classify the different methods. Yet it is interesting to note that glottal source features do not specifically appear in either work. It could be inferred that this kind of features are subsumed within the broader category 'voice quality', which only appears in Gold and French (2011). However, what the authors mean by voice quality is not actually explained in the article[3]. As a matter of fact, the definition of this parameter is not absent of complexity and ambiguity, as Gil and San Segundo (2014) tried to show. This concept is most frequently associated with perceptual analyses, mainly following the phonetic description of voice quality and the perceptual protocol described in Laver *et al.* (1981), known as Vocal Profile Analysis (VPA). The investigation of Stevens and French (2012) represents an example of the application of the VPA scheme to the characterization of voices for forensic purposes. While this protocol tries to objectify voice quality and it actually includes analysis categories related to the voice source (i.e. laryngeal tension, larynx position and phonation types), it remains a perceptual evaluation. The search for acoustical correlates of those perceptual measures is yet open to further investigation. The importance of undertaking this kind of research was already mentioned by Nolan (1983).

> It will be worth developing and improving this work [the work of specifying the acoustic correlates of an auditory phonetic framework for classifying voice qualities] since, from the point of view of speaker identification it provides an approach to the problem of classifying voices alternative, and complementary, to the more usual one of picking readily measurable acoustic features and investigating, in a relatively unguided way, how these features vary among a population of speakers. (Nolan, 1983: 108)

Taking into account the distinction (e.g. Jessen, 1997) between supralaryngeal voice quality and laryngeal voice quality, if we focus on the latter (i.e. voice aspects related to the glottal source), some forensic studies have aimed to investigate the speaker discriminatory potential of this type of features, from classical distortion parameters like jitter and shimmer (Künzel and Köster, 1992) to other laryngeal parameters related to the ratio between harmonics (Jessen, 1997), or later approaches suggesting the use of vocal source information to improve speaker recognition systems (Zheng, 2005). In this line, studies like Gómez-Vilda *et al.* (2008, 2009) or Gómez-Vilda *et al.* (2012) have proved that their voice analysis methodology — based on previous voice pathology investigations such as Gómez-Vilda *et al.* (2007) — is also useful for forensic speaker comparison. San Segundo and Gómez-Vilda (2013) or San Segundo and Gómez-Vilda (2014) represent some preliminary studies that have specifically tested in twins this methodology, which presents the advantage of splitting vocal from glottal information –by means of inverse filtering — thus opening the possibility of independently studying vocal and glottal components.

Research hypothesis and methodology

We start from the premise that a parameter that is genetically influenced will be a robust parameter for FSC. In other words, it will be highly speaker-discriminant for the comparison of the unknown and known speech samples. It is widely known in this discipline that some criteria exist for selecting a useful or robust forensic-phonetic parameter. Wolf (1972) set out these criteria and since then, other authors such as Nolan (1983) have spread and also redefined them. The first criterion (*high between-speaker variability*) and the second one (*low within-speaker variability*) are probably the most important, or at least they have been the most repeated criteria in many publications thereafter. These two criteria could be reformulated as: "the parameter needs to exhibit a high degree of variation from one speaker to another" (Nolan, 1983: 1) and "it should be as consistent as possible for each speaker" (Wolf, 1972: 2044). It seems logical to think that a parameter which is very dependent on the genetic endowment of the speaker will fulfill these two criteria.

For the purpose of evaluating whether a voice parameter is more or less 'genetic', San Segundo (2014) suggested the hypothesis that higher similarity values would be found in the comparison of MZ twin pairs than in DZ twin pairs, in pairs of non-twin siblings (in this case, male siblings, i.e. brothers) or in a population of unrelated speakers. This hypothesis applied to the three different analyses carried out in that study (aimed at investigating not only glottal source features but also formant trajectories of vocalic sequences and cepstral features). Since the current study focuses on glottal source features, our hypothesis would be as follows: *Glottal parameters will be genetically related: higher similarity values will be found in MZ twins than in DZ twins. These, in turn, will obtain higher similarity values than brothers (B), who will obtain higher similarity values than unrelated speakers (US).* The expected decreasing scale of similarity values in these speakers would then be: MZ $>$ DZ $>$ B $>$ US[4] According to this, we can establish the five following hypotheses:

H1. Intra-speaker comparisons should yield large likelihood ratios (LRs).
H2. MZ intra-pair comparisons should yield also large LRs.
H3. DZ intra-pair comparisons should yield large LRs although not as large as H1 or H2.
H4. B intra-pair comparisons should yield LRs at least over the background baseline.
H5. US intra-pair comparisons should yield LRs aligned with the background baseline.

Taking into account that the results of the speaker comparisons will be shown in the form of log-likelihood ratios (LLRs), the decision thresholds (λ) for the hypotheses described above could be represented as:

H1.$\lambda >$ -1
H2.$\lambda >$ -1
H3.$\lambda >$ -10
H4.$\lambda >$ -10
H5.$\lambda <$ -10

For the execution of this study, we have recruited 54 male speakers, distributed in four different groups:

- Monozygotic twins (MZ), also called identical twins: 24 speakers.
- Dizygotic twins (DZ), also called non-identical or fraternal twins: 10 speakers.

- Full brothers (B), i.e. of the same mother and same father: 8 speakers.
- Unrelated speakers (US), who for the most part were pairs of friends or work colleagues: 12 speakers.

Friends or work colleagues — the fourth speaker group — served to create a reference population, whose relevance for Likelihood-Ratio-based forensic comparison has been acknowledged elsewhere (e.g. Morrison, 2010). In short, a reference population is aimed at considering *typicality* in addition to *similarity*[5]. The ages of all the speakers recruited for this study ranged between 18 and 52 years old (median age: 28.96). The age difference between the siblings in each pair varied between four and eleven years. The language variety spoken by all the subjects was North-Central Peninsular Spanish. Speakers were recorded on two different recording sessions (separated by 2-4 weeks) in order to account for intra-speaker variability.

Although the participating speakers were recorded carrying out five different speaking tasks — for a full description of the *ad hoc* collected corpus, see San Segundo (2013) and San Segundo (2014) — in the current study we have specifically extracted the speech material from the fifth speaking task: informal interview with the researcher. This task was carried out on the telephone in the following way: the researcher is at one end of the telephone and one member of each speaker pair at a time is at the other end of the telephone[6]. In this task, which lasts around 5-10 minutes, the researcher asks the speaker about any of the topics[7] that they had been discussing with their conversational partner — either his sibling or friend — in the first task (semi-structured spontaneous conversation). Since there is a considerably long time gap between the execution of the first and the fifth task, the speakers do not remember clearly the whole conversation and they exhibit hesitating responses, resulting in *pause fillers* (Cicres, 2007).

The complete speech material consisted of 853 tokens of the [e:] vowel (average tokens per speaker and session: 7.89) naturally sustained in pause fillers. For the selection of the sustained [e:] vowels we made an auditory and spectrographic examination in *Praat* (Boersma and Weenink, 2012) for every speaker and session's audio files recorded in the fifth task. We did not select those vowels where we perceived a marked creak realization, a high degree of nasalization, overlap with extraneous noise, laughter, etc. In average, the duration of the vowels was around 200 milliseconds. These phonetic units were manually located with *Praat* and the most stable part of them was marked and extracted, avoiding the beginning and the end of the vowel. These pause fillers or hesitation marks, which most people use — as the name suggests — when they hesitate in a conversation, while they are thinking of what they are going to say next, or when they are trying to remember something, were found very useful for our study, as they are longer than vowels in connected speech. Obtaining a relatively long vowel is highly important in order to estimate glottal parameters, which in the clinical tradition have been normally elicited upon asking the subject to sustain a long vowel for as long as possible. This technique — which is foreign to the forensic realm — could be replaced by the use of naturally sustained pause fillers.

Using the software BioMet®Soft (2010), a vector of 68 parameters was created from each vocalic segment. These parameters were estimated from the glottal source by inverse filtering (Gómez-Vilda *et al.*, 2009) and they can be distributed in the following seven subgroups: 1) f0 and distortion parameters; 2) cepstral coefficients of the glottal source power spectral density (PSD); 3) singularities of the glottal source PSD; 4) biome-

chanical estimates of vocal fold mass, tension and losses; 5) time-based glottal source coefficients; 6) glottal gap (closure) coefficients; and 7) tremor (cyclic) coefficients. For a detailed description of these parameters, see San Segundo (2014). BioMet®Soft (2010) was also used to carry out the speaker comparisons in the form of pairwise parameter matching experiments, yielding the results in LRs, as in Ariyaeeinia *et al.* (2008). The specific methodology is described in Gómez-Vilda *et al.* (2012).

The vector of glottal features will be referred as x_{sij}, where *s* refers to the speaker, *i* is for the session, and *j* for the vowel filler. The two voice samples under test — in each comparison — will be denoted by Z_a=$\{x_{aij}\}$ and Z_b=$\{x_{bij}\}$ for subjects *a* and *b*. Thus, we will be evaluating our two-hypotheses contrasts in terms of a logarithmic likelihood value:

$$\lambda_{ab} = log\left(\frac{p(Z_b|\Gamma_a)}{\sqrt{p(Z_a|\Gamma_R)p(Z_b|\Gamma_R)}}\right)$$

where we evaluate the conditional probability of each speaker relative to a Reference Speaker's Model Γ_R and calculate if the conditional probability between the two voice samples *a* and *b* is larger than the conditional probabilities relative to the reference model. The conditional probabilities have been evaluated using Gaussian Mixture Models (Γ_a, Γ_b, Γ_R) as:

$$p(Z_b|\Gamma_a) = \Gamma_a(Z_b)$$

$$p(Z_a|\Gamma_R) = \Gamma_R(Z_a)$$

$$p(Z_b|\Gamma_R) = \Gamma_R(Z_b)$$

The forensic-comparison evaluation framework used is a two-step process, which could be described as follows:

- Step 1. Model Generation: A model representative of the reference population (male subjects between 18-52 years old) was created using recordings $\mathbf{Z}_R$=$\{x_{Rjk}\}$ as a Gaussian Mixture Model Γ_R=$\{w_R, \mu_R, C_R\}$ where w_R, μ_R and C_R are the set of weights, averages and covariance matrices, respectively, associated to each Gaussian Probability Distribution in the set.
- Step 2. Score Evaluation: The material under evaluation is composed of different parameterized voice samples grouped in a matrix $\mathbf{Z}_a$=$\{\mathbf{x}_{aj}\}$ where $1 \leq j \leq J_a$ is the sample index, each sample being a vector $\mathbf{x}_{aj}$=$\{\mathbf{x}_{aj1} \ldots \mathbf{x}_{ajM}\}$ from vowel segments conveniently parameterized. Similarly, the set of the corresponding speaker to be matched will be given as $\mathbf{Z}_b$=$\{\mathbf{x}_{bj}\}$ where $1 \leq j \leq J_b$ will be the sample index, each sample being a vector $\mathbf{x}_{bj}$=$\{\mathbf{x}_{bj1} \ldots x_{bjM}\}$. The conditioned probability of a sample from speaker *a* x_{aj} matching speaker *b* will be estimated as

$$P(x_{bj}|\Gamma_a) = \frac{1}{(2\pi)^{M/2}|C_a|^Q} \cdot e^{-1/2(x_{bj} - \mu_a)^T C_s^{-1}(x_{bj} - \mu_a)}$$

Similarly the conditioned probability of a sample from speaker a matching the Reference Model will be:

$$P\left(x_{aj}|\Gamma_R\right) = \frac{1}{(2\pi)^{M/2}|C_R|^Q} \cdot e^{-1/2(x_{aj} - \mu_R)^T C_s^{-1}(x_{aj} - \mu_R)}$$

Finally, the conditioned probability of a sample from speaker b matching the Reference Model will be:

$$P\left(x_{bj}|\Gamma_R\right) = \frac{1}{(2\pi)^{M/2}|C_R|^Q} \cdot e^{-1/2(x_{bj} - \mu_R)^T C_s^{-1}(x_{bj} - \mu_R)}$$

Results

The results of the different comparison tests are shown in tables 1 to 4, where we have marked whether the LLR values of each comparison entail the confirmation or the refutation of the hypotheses described in the previous section.

MZ speakers			
Speakers compared	**Type of comparison**	**LLR**	**Hypothesis confirmation**
01v01	Intra-speaker	2.4	√
02v02	Intra-speaker	-0.5	√
01v02	Intra-pair	-0.0	√
03v03	Intra-speaker	-1.1	×
04v04	Intra-speaker	-8.3	×
03v04	Intra-pair	-1.0	√
05v05	Intra-speaker	12.5	√
06v06	Intra-speaker	6.1	√
05v06	Intra-pair	5.8	√
07v07	Intra-speaker	12.0	√
08v08	Intra-speaker	6.6	√
07v08	Intra-pair	12.1	√
09v09	Intra-speaker	-7.0	×
10v10	Intra-speaker	23.0	√
09v10	Intra-pair	12.6	√
11v11	Intra-speaker	4.3	√
12v12	Intra-speaker	14.1	√
11v12	Intra-pair	-14.6	×
33v33	Intra-speaker	-5.0	×
34v34	Intra-speaker	0.2	√
33v34	Intra-pair	0.6	√
35v35	Intra-speaker	-1.6	×
36v36	Intra-speaker	-0.2	√
35v36	Intra-pair	-1.5	×
37v37	Intra-speaker	-7.0	×
38v38	Intra-speaker	15.7	√
37v38	Intra-pair	9.9	√
39v39	Intra-speaker	3.1	√
40v40	Intra-speaker	4.9	√
39v40	Intra-pai	2.9	√
41v41	Intra-speaker	6.9	√
42v42	Intra-speaker	-4.1	×
41v42	Intra-pair	0.2	√
43v43	Intra-speaker	0-0	√
44v44	Intra-speaker	3.0	√
43v44	Intra-pair	-0.1	√

Table 1. Results for the MZ speakers
LLR means log-likelihood ratio.

DZ speakers			
Speakers compared	**Type of comparison**	**LLR**	**Hypothesis confirmation**
13v13	Intra-speaker	6.4	√
14v14	Intra-speaker	-0.7	√
13v14	Intra-pair	1.7	√
15v15	Intra-speaker	-8.7	×
16v16	Intra-speaker	5.2	×
15v16	Intra-pair	-3.2	√
17v17	Intra-speaker	1.6	√
18v18	Intra-speaker	4.3	√
17v18	Intra-pair	-10.1	√
19v19	Intra-speaker	0.6	√
20v20	Intra-speaker	-7.7	√
19v20	Intra-pair	-0.4	√
45v45	Intra-speaker	-1.0	×
46v46	Intra-speaker	0.0	√
45v46	Intra-pair	3.4	√

Table 2. Results for the DZ speakers
LLR means log-likelihood ratio.

Non-twin brothers (B)			
Speakers compared	**Type of comparison**	**LLR**	**Hypothesis confirmation**
21v21	Intra-speaker	6.4	√
22v22	Intra-speaker	-0.7	√
21v22	Intra-pair	1.7	√
23v23	Intra-speaker	-8.7	√
24v24	Intra-speaker	5.2	√
23v24	Intra-pair	-3.2	√
47v47	Intra-speaker	1.6	√
48v48	Intra-speaker	4.3	×
47v48	Intra-pair	-10.1	√
49v49	Intra-speaker	0.6	×
50v50	Intra-speaker	-7.7	×
49v50	Intra-pair	-0.4	√

Table 3. Results for the B speakers
LLR means log-likelihood ratio.

Unrelated Speakers (US)			
Speakers compared	**Type of comparison**	**LLR**	**Hypothesis confirmation**
25v25	Intra-speaker	-42.2	×
26v26	Intra-speaker	-0.7	√
25v26	Intra-pair	-11.2	√
27v27	Intra-speaker	10.2	√
28v28	Intra-speaker	11.9	√
27v28	Intra-pair	-9.7	×
29v29	Intra-speaker	-0.2	√
30v30	Intra-speaker	7.5	√
29v30	Intra-pair	-13.2	√
31v31	Intra-speaker	6.1	√
32v32	Intra-speaker	5.2	√
31v32	Intra-pair	-12.7	√
51v51	Intra-speaker	-4.9	×
52v52	Intra-speaker	4.9	√
51v52	Intra-pair	-10.4	√
53v53	Intra-speaker	8.1	√
54v54	Intra-speaker	5.7	√
53v54	Intra-pair	-12.1	√

Table 4. Results for the US speakers
LLR means log-likelihood ratio.

In relation to H1, we have computed all the cases of intra-speaker dissimilarity in the four tables, and we have found that five out of the total 54 participating speakers seem to be in the limit of the established threshold (subjects 03, 35, 48, 49 and 50) while eight speakers show strong intra-speaker dissimilarity (subjects 04, 09, 15, 20, 33, 37, 42 and 51), and only one shows very strong dissimilarity (subject 25). Therefore, 14 out of 54 do not fulfil H1. However, since five speakers out of 54 obtain values very close to the established threshold, we could speak of 9 out of 54 speakers not fulfilling the hypothesis of intra-speaker similarity.

Regarding H2, we find two out of 12 pairs not fulfilling it (MZ pairs 11-12 and 35-36). The third hypothesis is not fulfilled in one out of five pairs (DZ pair 17-18), while H4 — which refers to non-twin siblings — is fulfilled in all four cases. Finally, only one pair of unrelated speakers is slightly over the baseline (speakers 27-28) out of 5 cases fulfilling H5. Therefore, in view of the results, the degree of hypothesis corroboration could be summarized as:

H1: 40/54; a relaxed threshold would be 45/54 = 83.3%
H2: 10/12 = 83.3%
H3: 4/5 = 80%
H4: 4/4 = 100%
H5: 5/6 = 83%

We will present our comparison results by means of a Tippett plot, since this is a standard graphical method for representing the LR results of a forensic comparison

system as well as a method for the evaluation of a system performance. As recalled in San Segundo (2014: 106), "this type of representation was proposed by Evett and Buckleton (1996) in the field of DNA analysis and it owes its name to the work of Tippett *et al.* (1968), who first referred to the concepts of 'within-source comparison' and 'between-source comparison' (cf Drygajlo *et al.*, 2003)". In this type of graph, two types of curves are displayed, each one representing the probability for one of the competing hypothesis: *Hp* or *Hd*. Typically the hypothesis of the prosecution (*Hp*) is that the offender and the suspect samples come from the same speaker, while the hypothesis of the defense (*Hd*) is that they belong to different speakers. However, for the speaker types that we are testing (MZ, DZ, B or US), our Tippett plot needs to be based on a more specific *Hd*. In other words, the hypothesis of the defense is not simply that the voice samples belong to different speakers but — depending on the type of speakers compared at each time — that the voice samples belong to either (a) MZ twins, (b) DZ twins, (c) non-twin siblings, or (d) unrelated speakers. For that reason, figure 1 shows only one line rising to the right (the black line), representing the cumulative distribution of LLRs for all the intra-speaker comparisons — *targets* in Automatic Speaker Recognition (ASR) terminology — while there are four different lines rising to the left (red, magenta, cyan and blue), each one representing a different type of intra-pair comparison (a-d), depending on the type of kinship relationship between the speakers being compared. These cases of intra-pair comparisons are also inter-speaker comparisons *sensu stricto* and they would be named *non-targets* in ASR terminology.

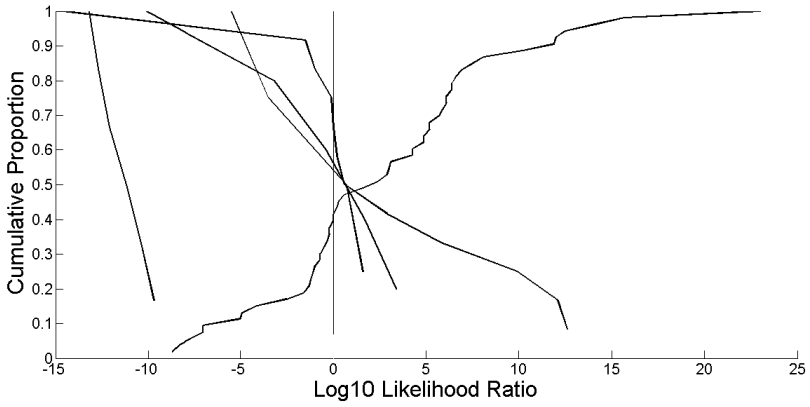

Figure 1. Tippett plot.
The black line represents the intra-speaker comparisons (for all the speaker types) and the following colours represent the intra-pair comparisons: red for US, magenta for MZ, cyan for DZ and blue for B.

As it can be seen in figure 1, the black line (intra-speaker comparisons) extends largely on the right, which implies a good performance of the system, but there are still some LLRs which support the contrary-to-fact hypothesis, represented in the black line from 0 to the left[8]. If we look at the intra-pair comparisons, different results are found:

- For the US (red line), the system performance is optimal, as there are only LLRs supporting the consistent-with-fact hypothesis. Note that all cases fall within the field to the left of 0. More specifically, the LLR values seem to be grouped

around -10, as could be also observed in table (cf. intra-pair comparisons). This indicates a *very strong support*[9] to the different-speaker hypothesis.

- In the case of MZ, DZ and B comparisons, the following trends are observed: The strongest support for the contrary-to-fact hypothesis occurs in MZ twins. Note that the magenta line stretching from 0 to the right is the longest. However, for the DZ (cyan) and B (blue) comparisons, the system shows a similar performance, with most cases falling within the consistent-with-fact hypothesis and only some cases supporting the contrary-to-fact hypothesis.

Discussion of the results

Our main hypothesis was that the glottal parameters analyzed would be genetically influenced, i.e. higher similarity values would be found in MZ twins than in DZ twins, non-twin brothers or in the reference population. Therefore, we predicted a decreasing scale of similarity values, expected to follow this order: MZ > DZ > B > US. According to this, we suggested five specific hypotheses, depending on whether the comparisons were intra-speaker comparisons (H1), or intra-pair comparisons of some of these types: MZ intra-pair comparison (H2), DZ intra-pair comparison (H3), B intra-pair comparison (H4) and US intra-pair comparison (H5). We further established some decision thresholds for each of these hypotheses in order to assess whether the LLR values obtained in the comparisons could be deemed large or small — and could consequently allow the rejection or the confirmation of the hypotheses.

In view of the results, the degree of hypotheses corroboration was very high: three of our hypotheses were corroborated in 83.3% of the cases (H1, H2 and H5), another one was corroborated in 80% of the cases (H3) and a further one was corroborated in 100% of the cases under study (H4). In the rest of this section we aim to discuss these results, distinguishing between the intra-speaker comparisons (which relate to H1) and the inter-speaker comparisons, referring to H2, H3, H4 and H5.

Intra-speaker results

There are no clear reasons why 14 out of 54 intra-speaker comparisons (or 9 out of 54 if we relax the threshold, as explained above) yield very low LLRs, indicating a strong dissimilarity of those speakers towards themselves. For the intra-speaker comparisons, the vowel fillers extracted from the first recording session are tested against the vowel fillers obtained in the second recording session. Some possible explanations for the cases of hypothesis rejection could then be associated with changes in phonation due to emotional stress or with the existence of temporary pathological conditions. Despite the fact that the speakers were only recorded when they exhibited a healthy condition — and the health troubles potentially affecting their voice had to be indicated by the speakers in a questionnaire —, it is still possible that they could have experienced temporary and minor voice maladies at one recording session but not at the other, this being behind the dissimilarity results of certain speakers.

Another possible explanation could be related with the speaker classification first proposed by Doddington *et al.* (1998). It is a truism in speaker recognition that not all speakers affect the performance of a forensic-comparison system in the same way, or in the words of the above-mentioned authors, there are "striking performance inhomogeneities among speakers within a population" (Doddington *et al.*, 1998: 1). The existence of these inhomogeneities allowed the authors to classify speakers in *sheep*,

lambs, *wolves* and *goats*; basically depending on whether they are more or less difficult to recognize by the system. In this sense, the percentage of speakers in our study who obtained low LLRs could probably be considered 'goats' in Doddington's Zoo, as these type of speakers "tend to adversely affect the performance of systems by accounting for a disproportionate share of the missed detections" (Doddington *et al.*, 1998: 1). Yet, independently of the fact that the existence of 'goats' is acknowledged since long in ASR, the question of what makes a speaker so different from himself is a key issue in Forensic Phonetics and it remains largely unexplored.

Finally, a distinction should be made between the average LLR values of the discordant cases of intra-speaker comparisons (around -8, -7 or lower) and a single case with a striking LLR value of -42.2 (speaker 25). As the study by San Segundo (2014) explained, this was a clear exception which deserved detailed analysis. Indeed, upon examination of the anamnesis of this speaker, it became apparent that he suffered from hypothyroidism. We suggested in the above-mentioned study that this hormonal problem could be the cause of the strikingly large intra-speaker variation found for this speaker. Often called underactive thyroid hormone, one of its symptoms is hoarse voice, according to Longo and Fauci (2011). This type of phonation, especially if it appears intermittently in the speaker's vocal output, could explain the strong dissimilarity in the comparison of the first and the second recording sessions of speaker 25. Nevertheless, more research would be necessary to investigate how this disease specifically affects voice.

Intra-pair (inter-speaker) results

Having focused on H1 in the previous subsection, we have to consider now separately H2, H3, H4 and H5. As far as H2 is concerned, only 2 out of 12 MZ pairs did not obtain LLRs above -1, as our hypothesis established. While one case is that of MZ pair 11-12 (LLR = -14.6), with a strong deviation from the established threshold, the other case is that of MZ pair 35-36 (LLR = -1.5), i.e. certainly close to the threshold. It seems evident that their cases are not comparable and that the most interesting pair to examine in detail is the first one. The most plausible reason for their striking differences — despite being identical twins — is twofold. On the one hand, the existence of smoking habits in one of them made his f0 much lower than that of his cotwin, and this could affect the rest of the glottal parameters analyzed in this study. On the other hand, the questionnaire that the speakers had to fill at the time of the recordings included some questions about their attitude towards being twins. In view of the answers given by this specific pair, it was made clear that they were not especially close to each other, which could have made them separate in personality and possibly also phonetically [10]. In other words, "the learned speech habits aimed at attaining divergence patterns may have outweighed their anatomical similarities" (San Segundo, 2014: 188).

As far as H3 is concerned, only in one DZ pair out of five the hypothesis was not corroborated. It is the case of DZ pair 17-18, who obtained a LLR = -10.1. In our hypothesis formulation, we considered that DZ twins should show large LLRs but not as large as MZ twins, being the decision threshold λ = -10. The only exception found is therefore almost irrelevant. In all the other cases, the LLR values were as expected: relatively large but not that large as those found for MZ twins, on average. For that reason, the third hypothesis is well corroborated.

If we consider now H4, all the non-twin brothers corroborate our hypothesis: LLR values above -10 are obtained in 100% of the pairs analyzed. Since full siblings (i.e. broth-

ers) and DZ twins both share the same genetic load, H3 and H4 were established at the same level: λ = -10. Finally, H5 established that US would obtain LLRs aligned with a background baseline fixed at λ < -10. This is fulfilled in almost all the cases, being the only exception that found in speakers 27-28 (LLR = -9.7). While this value implies a rejection of the hypothesis if we strictly apply our decision threshold, it seems clear that the difference between -9.7 and -10 is almost irrelevant, especially when we are expressing the results in logarithmic figures. The degree of H5 corroboration is then very satisfactory, and this is particularly relevant, as it indicates that in a typical forensic scenario — when unrelated speakers are compared — our glottal source based system performs very well, with none of the speakers being misidentified (false alarms). Besides, with these results more evidence is gained in favor of our main hypothesis that glottal parameters are genetically influenced, as none of the unrelated speakers show any similarity, in comparison with the somehow genetically related DZ and B — with larger LLR values — and with the much genetically related MZ, with still larger LLR values.

Conclusions and directions for future research

We can conclude that the glottal parameters analyzed, considered as a whole set of 68 features, are genetically influenced. With few exceptions, the system performance for DZ and full siblings is similar (λ > -10) while MZ twins obtain larger LLRs and the values of US gather homogenously around the baseline (λ < -10). This is in agreement with our hypotheses, as we predicted that the LLR values of the forensic comparison would be distributed in a line going from the largest positive LLRs for the MZ twins, at one end of the line, and the largest negative LLRs for the US, at the other end of the line. The former share 100% of their genes while the latter share 0%. In between, there are the DZ twins and the B, sharing on average 50% of their genetic information. Furthermore, our results are in agreement with previous studies about twins, such as Loakes (2006), insofar as different results have been found for different twin pairs, indicating a lack of homogeneity in this speaker group. The idiosyncrasies in the relationship of each pair could be only studied on a case-by-case basis to find the causes for speech convergence or divergence, which probably indicates that the weight of external factors, such as psychological aspects, or educational and environmental influences (i.e. 'nurture') is more important than it could be a priori thought in this type of voice studies, or at least as important as 'nature' in many speaker comparisons.

All in all, this study has tried to show the relevance of applying the twin methodology to forensic voice investigations in order to find whether a parameter — or a set of parameters, as in this case — could be robust and hence useful for speaker comparison. It becomes also apparent that the study of glottal source features deserves an important position among the many possible phonetic parameters that can be considered in forensic casework, both for their easy extraction in natural speech and for their good discrimination results in several studies so far, not to mention that they are obtained from inverse filtering of the vocal tract. This makes them independent from traditional vocal-tract features, which opens great possibilities for their combination with such parameters; this fusion/combination being one of the advantages of the LR approaches.

Besides studying the degree of similarity in MZ, DZ, B and US intra-pair comparisons, we have also taken advantage to study intra-speaker variation in all the four types of speakers participating in this study. We have found that in more cases than desirable in a forensic context, the system performance is not completely good when two speaker

sessions are tested against each other. These *missed hits* represent a 16.6% of the intra-speaker comparisons, or targets, regardless of the fact that the speaker is MZ, DZ, B or US. There was an especially striking case of LLR = -42.2. While this is a clear exception, it was in-depth analyzed and a possible explanation for this large intra-speaker variation could be found in a hormonal disease suffered by this speaker. Yet the other cases of missed hits still represent large figures whose cause would deserve further research. Likewise, future studies could consider the study of the 68 glottal parameters independently, to test if some of the seven feature subgroups outperform the others for forensic purposes.

Acknowledgments

This research has been possible thanks to a grant awarded by the Spanish Ministry of Education (*Beca FPU – Programa Nacional de Formación de Profesorado Universitario*; reference AP2008-01524) and also thanks to a grant awarded by the IAFPA (International Association for Forensic Phonetics and Acoustics).

Notes

[1]Some terminology controversies have arisen in recent times in relation to the proper name that this specific application of Forensic Phonetics should receive. It seems that the term comparison is widespread nowadays, at least in the linguistic-phonetic realm. The Position Statement in French and Harrison (2007), signed by nine researchers and with several more co-signatories, accepts the replacement of identification by comparison: "It will be apparent from the arguments developed here that the term FSI should be replaced by FSC" (French and Harrison, 2007: 144). A summary of this controversy can be read in San Segundo (2014). For more details, see Coulthard and Johnson (2007); French and Harrison (2007); Morrison (2009); Rose and Morrison (2009) and French *et al.* (2010).

[2]Note however that *speaker profiling* does not necessarily involve LADO. The former simply consists in determining the phonetic profile of an unknown speaker on the basis of his voice and speech patterns; i.e. trying to derive as much information as possible about the speaker age, gender or dialect, among other characteristics.

[3]Most probably because providing a definition of 'voice quality' is clearly not the purpose of the investigation by Gold and French (2011). In fact, this is not an easy concept to define. In Gil and San Segundo (2014) we track the description of 'voice quality' in six of the most relevant works about Forensic Phonetics — including references to voice quality — to this date (Gil and San Segundo, 2014: 176-183). Namely, the reviewed works were, in chronological order: Nolan (1983), Hollien (1990), Künzel (1994), French (1994), Rose (2002) and Jessen (2008). An examination of those works allows the reader to see how far speech scientists are from arriving at a definition consensus. No wonder Hollien (1990) points to the occasional view of the label 'voice quality' as a 'wastebasket' used for those voice aspects that other categories fail to describe.

[4]If we strictly apply what we know about the genetic endowment of DZ and B (as explained above: same amount of shared genes per sibling pair, that is 50%), it could be thought that it would have been more coherent to establish this decreasing scale MZ > DZ ≥ B > US. Yet, two aspects should be taken into account: a) Although it is widely accepted that both DZ pairs and non-twin sibling pairs "share 50% of their genes, on average, by descent" (Pakstis *et al.*, 1972 in Segal, 1990: 612), a more realistic range seems to be 25% - 75% while this theoretical range can actually vary between 0% to 100% (Pakstis *et al.*, 1972 in Segal, 1990: 612). Therefore it should be highlighted that the 50% value is — to some degree – a convention; it can vary from one pair to another. b) The newly-developed scientific field of epigenetics (the study of the changes in gene expression caused by mechanisms other than changes in the underlying DNA sequence) has shown us that environmental factors do affect genes in ways that still need to be fully explored. As environmental and genetic aspects cannot be completely disentangled, we consider that DZ twins could be –although maybe only slightly — more genetically related than non-twin siblings because the former usually share more environmental experiences than the latter due to the fact that they are born on the same day whereas in the case of non-twin siblings their age gap makes them more susceptible for

environmental divergence. There are several arguments supporting that DZ cotwins are genetically more similar to one another than non-twin siblings. The interested reader is encouraged to read — for instance — Stromswold (2006), where she raises the case of transplant surgery, a field where "it has been known for decades that the incidence of graft rejection is lower between DZ cotwins than between non-twin full siblings, and this clinical observation has been used to argue that DZ cotwins are genetically more similar to one another than non-twin full siblings (see Geschwind, 1983)" (Stromswold, 2006: 338–9). As all these genetic aspects are not free of controversy, it seems prudent for us to maintain the hypothesized decreasing scale MZ > DZ > B > US while — at the time of fixing the thresholds for the corresponding H3 and H4 — establishing lambda at the value -10 in both cases. This is not in contradiction with the explanation of each hypothesis: (H3) *DZ intra-pair comparisons should yield large LRs although not as large as H1 or H2*, and (H4) *B intra-pair comparisons should yield LRs at least over the background baseline.*

[5]The LR formula has a numerator and a denominator. As explained in Morrison (2010: 17), "the numerator of the LR can be considered a *similarity* term, and the denominator a *typicality* term. In calculating the strength of evidence, the forensic scientist must consider not only the degree of similarity between the samples, but also their degree of typicality with respect to the relevant population. In fictional television shows, forensic scientists are often portrayed comparing two objects, finding no measurable differences between them, and shouting: 'It's a match!' Similarity alone, however, does not lead to strong support for the same-origin hypothesis".

[6]This does not mean that the speech material available for comparison had been telephone-filtered. The corpus used contains some speaking tasks which have undergone a filtering through real telephone transmission, but on this occasion we used studio-quality recordings. The recording set-up was such that the speakers were at different rooms and held a real telephone conversation but they were being recorded with high quality microphones (*Countryman E6i Earset* microphone). The recordings took place in the *Centro de Ciencias Humanas y Sociales* at *CSIC (Consejo Superior de Investigaciones Científicas)* in Madrid, Spain.

[7]In the first speaking task, the speakers were suggested some topics, including those described in Loakes (2006). For especially sparing speakers, other possible topics were raised. In order to minimize the "observer's paradox" (Labov, 1972), we followed the indications in Moreno (2011), particularly with regard to the use of "icebreakers" as conversational starting points.

[8]Note that Speaker 25 (only the value for his intra-speaker comparison, i.e. LLR = -42) was excluded from representation in the black line of figure because that LLR value was considered an outlier, i.e. being exceptionally low for the reasons which will be more thoroughly discussed in the section devoted to the discussion of the results.

[9]Note that a LLR of -10 (LR = -10, 000, 000, 000) means that it is 10, 000, 000, 000 times more likely that the observed differences between the speech samples of suspect and offender occur under the hypothesis that they come from different speakers than under the hypothesis that they come from the same speaker. According to the verbal equivalents for LRs proposed by Evett (1998), LRs larger than 1000 indicate a *very strong support* for the respective hypothesis (in this case, the hypothesis of the defense, as it is a negative logarithmic value). Nevertheless, it should be noted that not all scientists agree in using such verbal scales (for a summary of this controversy, see San Segundo, 2014, *cf. Introduction).*

[10]In the questionnaire, they rated their relationship closeness as "not especially close" and answered that they have liked to be independent and different since they were children. This compares with the most common situation for the rest of MZ twins participating in this study, who — on average — rated their relationship closeness as "very close" and stated that they like to be together and share leisure activities, group of friends, etc.

References

Ariyaeeinia, A., Morrison, C., Malegaonkar, A. and Black, S. (2008). A test of the effectiveness of speaker verification for differentiating between identical twins. *Science & Justice*, 48(4), 182–186.

BioMet®Soft, (2010). Biomet®soft. universidad politécnica de madrid. retrieved from httpÂů//www.biometrosoft.com.

Boersma, P. and Weenink, D. (2012). Praat: doing phonetics by computer (version 5.3.79).

Cambier-Langeveld, T. (2007). Current methods in forensic speaker identification: Results of a collaborative exercise. *International Journal of Speech, Language and the Law*, 14(2), 223–243.

Cicres, J. (2007). Análisis discriminante de un conjunto de parámetros fonético-acústicos de las pausas llenas para identificar hablantes. *Síntesis Tecnológica*, 3(2), 87–98.

Coulthard, M. and Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. New York: Routledge.

Debruyne, F., Decoster, W., Van Gijsel, A. and Vercammen, J. (2002). Speaking fundamental frequency in monozygotic and dizygotic twins. *Journal of Voice*, 16(4), 466–471.

Doddington, G., Liggett, W., Martin, A., Przybocki, M. and Reynolds, D. (1998). Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proceedings of the 5th International Conference on Spoken Language Processing*, 1–5.

Drygajlo, A., Meuwly, D. and Alexander, A. (2003). Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, 689–692.

Evett, I. (1998). Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice*, 38(3), 198–202.

Evett, I. and Buckleton, J. (1996). Statistical analysis of str data. In A. Carraredo, B. Brinkmann and E. Bär, Eds., *Advances in Forensic Haemogenetics*. Heidelberg: Springer-Verlag.

French, P. (1994). An overview of forensic phonetics with particular reference to speaker identification. *International Journal of Speech, Language and the Law*, 1(2), 169–181.

French, P. and Harrison, P. (2007). Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French & Philip Harrison. *The International Journal of Speech, Language and the Law*, 14(1), 137–144.

French, P., Nolan, F., Foulkes, P., Harrison, P. and McDougall, K. (2010). The UK position statement on forensic speaker comparison; a rejoinder to Rose and Morrison. *The International Journal of Speech, Language and the Law*, 17(1), 143–152.

French, P. and Stevens, L. (2013). Forensic speech science. In M. Jones and R. Knight, Eds., *The Bloomsbury Companion to Phonetics*. London: Bloomsbury.

Geschwind, N. (1983). Genetics: fate, chance, and environmental control. In C. L. . J. Cooper, Ed., *Genetics aspects of speech and language disorders*, 21–33. New York: Academic Press, 1 ed.

Gil, J. and San Segundo, E. (2014). La cualidad de voz en fonética judicial. In E. Garayzábal and M. J. M. Reigosa, Eds., *Lingüística Forense: la lingüística en el ámbito legal y policial*. Madrid: Euphonia Ediciones.

Gold, E. and French, P. (2011). An international investigation of forensic speaker comparison practices. In *Proceedings of the 17th International Congress of Phonetic Sciences*, 1254–1257, Hong Kong, China.

Gómez-Vilda, P., Fernández-Baillo, R., Nieto, A., Díaz, F., Fernández-Camacho, F. J., Rodellar, V., Alvarez, A. and Martínez, R. (2007). Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters. *Journal of Voice*, 21(4), 450–476.

Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, M. V., Nieto-Lluis, V., Álvarez Marquina, A., Mazaira-Fernández, L. M. and Godino-Llorente, J. I. (2009). Glottal

source biometrical signature for voice pathology detection. *Speech Communication*, 51(9), 759–781.

Gómez-Vilda, P., Mazaira-Fernández, L. M., Martínez-Olalla, R., Álvarez Marquina, A., Hierro, J. and Nieto, R. (2012). Distance Metric in Forensic Voice Evidence Evaluation using Dysphonia-relevant Features. In *VI Jornadas de Reconocimiento Biométrico de Personas (JRBP)*, Las Palmas de Gran Canaria.

Gómez-Vilda, P., Álvarez Marquin, A., Mazaira-Fernández, L. M., Fernández-Baillo, R., Nieto-Lluis, V., Martínez-Olalla, R. and Rodellar-Biarge, M. V. (2008). Decoupling vocal tract from glottal source estimates in speaker's identification. *Language Design*, Special Issue, 111–118.

Gómez-Vilda, P., Álvarez Marquina, A., Mazaira-Fernández, L. M., Fernández-Baillo, R., Rodellar-Biarge, M. V. and Nieto-Lluis, V. (2010). Glottal biometric features: Are pathological voice studies appliable to voice biometry? In *I Workshop de Tecnologías Multibiométricas para la Identificación de Personas*, Las Palmas de Gran Canaria.

Hazen, K. (2002). The family. In J. Chambers, P. Trudgill and N. Schilling-Estes, Eds., *The Handbook of Language Variation and Change*. Malden, MA: Blackwell.

Hollien, H. (1990). *The acoustics of crime*. New York: Plenum Press.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *The International Journal of Speech, Language and the Law*, 4(1), 84–103.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2(4), 671–711.

Künzel, H. (1994). Current approaches to forensic speaker recognition. In *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 135–141.

Künzel, H. (2010). Automatic speaker recognition of identical twins. *The International Journal of Speech, Language and the Law*, 17(2), 251–277.

Künzel, H. and Köster, J. (1992). Measuring vocal jitter in forensic speaker recognition. In *Proceedings of the 44th Annual Meeting, American Academy of Forensic Sciences*, 113–114.

Labov, W. (1972). The transformation of experience in narrative syntax. In W. Labov, Ed., *Language in the Inner City*. Philadelphia: University of Philadelphia Press.

Laver, J., Wirz, S., Mackenzie, J. and Hiller, S. M. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress*, 14, 139–155.

Loakes, D. (2006). *A forensic phonetic investigation into the speech patterns of identical and non-identical twins*. Doctoral dissertation, University of Melbourne.

Longo, D. and Fauci, A. (2011). Disorders of the thyroid gland. In D. Longo, A. Fauci, D. Kasper, S. Hauser, J. Jameson and S. Loscalzo, Eds., *Harrison's Principles of Internal Medicine*. New York: McGraw-Hill.

Mora, M. (2013). La policía francesa detiene a dos gemelos para aclarar una ola de ataques sexuales. *El País*, Retrieved from http://internacional.elpais.com/internacional/2013/02/10/actualidad/1360530132_840599.html, 10 February.

Moreno, F. (2011). La entrevista sociolingüística: esquemas de perspectivas. *Linred: lingüística en la Red*, 9, 1–16.

Morrison, G. (2009). Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. *Australian Journal of Forensic Sciences*, 41(2), 155–161.

Morrison, G. (2010). Forensic voice comparison. In I. Freckelton and H. Selby, Eds., *Expert Evidence*. Sydney: Thomson Reuters.

Nolan, F. (1983). *The phonetic bases of speaker recognition.* Cambridge: Cambridge University Press.

Pakstis, A., Scarr-Salapatek, S., Elston, R. and Siervogel, R. (1972). Genetics contributions to morphological and behavioral similarities among sibs and dizygotic twins: Linkages and allelic differences. *Social Biology*, 19, 185–192.

Rose, P. (2002). *Forensic Speaker Identification.* London: Taylor & Francis.

Rose, P. (2006). Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, 20(2), 159–191.

Rose, P. and Morrison, G. S. (2009). A response to the UK position statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law*, 16(1), 139–163.

San Segundo, E. (2013). A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application. *Procedia-Social and Behavioral Sciences*, 95, 59–67.

San Segundo, E. (2014). *Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics.* Doctoral dissertation, CSIC/UIMP.

San Segundo, E. and Gómez-Vilda, P. (2013). Voice biometrical match of twin and non-twin siblings. In *Proceedings of the 8th International Workshop Models and analysis of vocal emissions for biomedical applications*, 253–256, Firenze, Italy.

San Segundo, E. and Gómez-Vilda, P. (2014). Forensic voice comparison using glottal parameters in twins and non-twin siblings. In *The 23rd Conference of the International Association for Forensic Phonetics and Acoustics*, Zürich, Switzerland.

Segal, N. (1990). The importance of twin studies for individual differences research. *Journal of Counseling & Development*, 68(6), 612–622.

Stevens, L. and French, P. (2012). Voice quality in Standard Southern British English: distribution of features, inter-speaker variability and the effect of telephone transmission. In *The 21st Conference of the International Association for Forensic Phonetics and Acoustics*, Santander, Spain.

Stromswold, K. (2006). Why aren't identical twins linguistically identical? Genetic, prenatal and postnatal factors. *Cognition*, 101(2), 333–384.

Tippett, C., Emerson, V., Fereday, M., Lawton, F., Richardson, A., Jones, L. and Lampert, M. (1968). The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society*, 8(2), 61–65.

Tomblin, J. and Buckwalter, P. (1998). Heritability of poor language achievement among twins. *Journal of Speech, Language, and Hearing Research*, 41(1), 188.

Wolf, J. (1972). Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, 51(6B), 2044–2056.

Zheng, N. (2005). *Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract.* Doctoral dissertation, The Chinese University of Hong Kong.